Original Instruction

"Walk down the corridor and upstairs. Stop halfway up the stairs."

Visual Observations

Temporal Reordering LLM:
- Sub-goals:
Walk down the corridor. Reach the stairs. Walk upstairs. Stop halfway up the stairs.

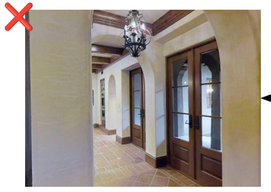
VLM-based Action Router:
- Previous sub-goals:
['Walk down the corridor']

- Sub-goal (current):
Reach the stairs

- Reasoning:
The current image shows the entrance to the house, and the next logical step is to move towards the stairs as instructed.

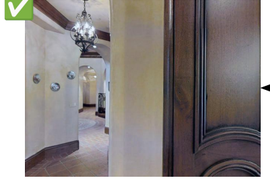
ScaleVLN Agent

Action Image



Vertical Movement Agent

Action Image



(a) A sample in R2R Val Unseen

Original Instruction

"Alright, so what you'll want to do is walk straight ahead, and then, um, take a left turn. Keep going forward until you reach the pillars, and, let's see, just wait there in the middle. It's hard to miss."

Visual Observations

Temporal Reordering LLM:
- Sub-goals:
Walk straight ahead. Turn left. Continue forward. Reach the pillars. Stop and wait in the middle of the pillars.

VLM-based Action Router:
- Previous sub-goals:
['Walk straight ahead', 'Turn left', 'Continue forward', 'Reach the pillars']

- Sub-goal (current):
Stop and wait in the middle of the pillars

- Reasoning:
The agent has reached the pillars and is positioned in the middle, as indicated by the previous images showing the agent approaching and then standing in the middle of the pillars.

ScaleVLN Agent

Action Image

Stop and Pause Agent

Action Image

(b) A sample in GSA-R2R Test-N-Scene